Chapter 2
Literature Review

This chapter first introduces the transmission channels employed currently for speech communication and their main impairments and then presents the literature review, divided into three parts: channel quality evaluation, human speaker recognition, and automatic speaker recognition. Different procedures for evaluation and main outcomes relevant to this work are indicated.

The review of channel quality evaluation reports the current status of investigations addressing subjective perceptions and automatic evaluations of signal quality when the speech is transmitted through different kinds of communication channels. The rest of this review shows state-of-the-art methods to assess the human and the automatic speaker recognition performances, and the channel impairment effects that have been reported in previous investigations. On the human side, pertinent listening tests to assess the human capability to detect speaker identities reveal how the performance is influenced by different voice distortions. On the automatic side, a review of the most recent and efficient methods for automatic speaker recognition and their main findings under channel degradations are presented.

Based on the fact that channels of extended bandwidths generally offer better quality and on the assessed importance of different speech frequency ranges for speaker recognition, this book concentrates on evaluating the advantages of enhanced channels for the human and for the automatic speaker recognition performance, clarifying how transmissions affect the speaker-specific voice properties and their relation to signal quality measurements.

2.1 Today's Communication Channels and Their Main Impairments

A communication channel is referred to in this book as an end-to-end physical medium with attached devices through which audio signals—only voice signals are of interest in this work—are transmitted. It involves thus electro-acoustic user

L. Fernández Gallardo, *Human and Automatic Speaker Recognition over Telecommunication Channels*, T-Labs Series in Telecommunication Services, DOI 10.1007/978-981-287-727-7_2

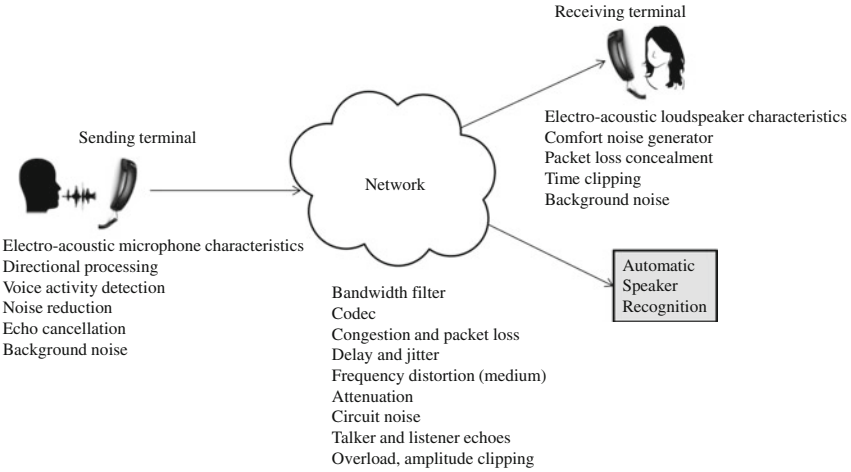

Fig. 2.1 General end-to-end communication channel and indication of its main impairments

interfaces at both ends of the communication (i.e. devices with a microphone and a loudspeaker in sending and in receiving direction, respectively) and the transmission medium itself (e.g. copper wire, wireless, fibre optic, etc.). Figure 2.1 shows a general communication channel and its main impairments, associated to the different channel components, which affect the transmitted audio signal.

With today's rapid deployment of digital transmissions and advances in technology, assorted communication networks are available. The Public Switched Telephone Network (PSTN) consists of different transmission mediums interconnected by switching centres, permitting users to place telephone calls to practically anywhere in the world, and is the primary platform for voice communications. It comprises the traditional landline telephony—partially replaced by the digital Integrated Services Digital Network (ISDN)—and mobile telephony such as the Global System for Mobile Communication (GSM). Alternatively, long-distance communication between two people is also enabled by packet switched networks, adapted for speech transmission employing the VoIP technology. The speech data are compressed and embedded in IP packets to be sent, discontinuously, over the Internet.

The capacity of the channel for transmitting information can be given in terms of its bandwidth, measured in Hz, or its bitrate, measured in bits per second. A passband filter is incorporated to remove high and low frequency components, fixing the channel bandwidth. A speech coding algorithm is then necessary to compress the data for an efficient transmission, that is, to reduce the transmission delays and to respect the channel bandwidth constraints. Speech codecs can generally operate at different bitrates, depending on the application requirements. Most of the current speech codecs can be broadly classified into waveform codecs, which aim at reconstructing the speech waveform; parametric codecs, which aim at reconstructing a new speech waveform on the basis of transmitted parameters (e.g. parameters from a speech production model); or hybrid codecs, which combine the previous two principles.

The technical operation of networks and codecs adheres to the standards created by the International Telecommunication Union, Telecommunication Standardization Sector (ITU-T) and by the European Telecommunications Standards Institute (ETSI). Among other regulations in telecommunications to assure interoperability, these standards deal with specific quality aspects and with guidelines for conducting evaluations of channel quality.

Because the majority of the energy of speech signals is concentrated between about 300 and 3 kHz and because channel bandwidth was very precious a few decades ago, the telephone bandwidth from 300 to 3.4 kHz was standardised, termed narrowband (NB). However, it was later demonstrated that an extended bandwidth, namely wideband (WB), from 50 to 7 kHz, provided better quality and intelligibility compared to NB. The sampling rate is increased to 16 kHz, approximately doubling the NB frequency range. The low frequencies incorporated contribute to increased naturalness, presence, and comfort, whereas the high frequency extension facilitates fricative differentiation [133]. These findings, together with cost reductions of network resources led to the standardisation of WB. Even more extended is the range of the emerging super-wideband (SWB) transmissions, 50–14,000 Hz, intended for high-quality videoconferencing, although not yet widely deployed.

The PSTN remains limited to NB, as originally designed, and typically employs the codec G.711 at 64 kbit/s, whereas VoIP services offer also WB using a broader variety of codec schemes. The most applied ones are: G.711, G.723.1, G.726, G.727, G.728, and G.729 in NB; and G.722 and Adaptive Multi-Rate (AMR)-WB, also termed G.722.2, in WB. Typical codecs employed in mobile networks are GSM-Half Rate (GSM-HR), GSM-Full Rate (GSM-FR), GSM-Enhanced Full Rate (GSM-EFR), and AMR-NB. Recent SWB codecs are AMR-WB$^+$, G.722.1C, G.718B, G.711.1D, G.722B, AAC-ELD, and Opus, to name a few. Another emerging codec for packet networks supporting NB, WB, and SWB is Speex. The choice of the codec depends on the target application and on a trade-offs between bitrate, quality, robustness, complexity or processing power, and delay. The coding-decoding processes modify the spectral characteristics of the original speech signal as they introduce undesirable non-linear distortions. Hence, together with the channel bandwidth filter, the different codec implementations affect the speech quality to different extents. The degradation is more accentuated in NB and with codecs operating at lower bitrates, compared to WB and to codecs with a lesser compression level. A detailed description of ITU and ETSI codecs employed for different bandwidths can be found in [49, 133, 263].

Despite the better quality offered by WB and SWB channels, most communications are still limited to NB due to the prevalence of the PSTN infrastructure. A call IP-to-PSTN, although may initially be of WB quality, is limited by the 3.1 kHz PSTN bandwidth. Besides, the user terminals with which the conversation is initiated or terminated do not always support WB and constrain the transmitted signal spectrum to the devices' frequency range. These facts have motivated research on the so-termed Artificial Bandwidth Extension (ABE), with the objective of enhancing the transmitted NB speech at the receiving side. Proposed methods to synthesise a WB signal, such as the commonly employed *envelope aliasing* or others based on a linear model

of the human speech production, result in a better quality of the reconstructed signal, yet a clear gap to WB speech still remains [257]. In particular, fricative sounds cannot be well estimated from a NB signal since most of their distinctive energy is concentrated above 4 kHz [14, 83]. Moreover, the speaker-specific characteristics are not normally synthesised in the artificially added frequency ranges, as the typical methods for ABE are trained on speech from multiple speakers [41].

Notwithstanding that it will require some time and effort until terminals and networks support the WB transmission, VoIP is expected to replace the PSTN in the near future. In contrast to the PSTN, initially designed and optimised for analogue transmissions, the VoIP technology offers not only the benefits of WB, but also higher flexibility and a cost reduction, due to a more efficient use of the bandwidth over the IP infrastructure. As drawbacks, VoIP is more complex and introduces different channel impairments (other than bandwidth filter and codec) such as One Way Delay (OWD), jitter, and packet loss, which may significantly affect the quality of VoIP [259]. A number of investigations have addressed the measurements of end-to-end delay and packet loss in different transmission configurations [259] as well as techniques to deliver Quality of Service (QoS) guarantees for the users [193]. Additional motivation for the transition from NB to WB communications is provided in this book, which shows the benefits of extended bandwidths for recognising speakers.

Packet loss, which occurs as a result of congestion in the network, may provoke severe voice quality degradations, being the impairment which makes VoIP perceptually most different from the circuit switched network [210]. The degradations can result in choppy, garbled or even unintelligible speech. Because of its time varying nature, the packet loss rate can be modelled as random, where a packet is lost with a certain probability, or bursty, which reflects better the real network congestions where losses may extend over several packets. Decoders may implement Packet Loss Concealment (PLC) methods by inserting silence, noise, or a reconstructed packet based on the speech signal in the neighbourhood of the lost packet or packets, alleviating the loss of quality to some extent.

The user interfaces employed in communication channels introduce further distortion in sending and in receiving direction, due to the intrinsic characteristics of their microphones and loudspeakers and their integration into the physical device. Most microphones found in telephony perform adequately in the 80–10,000 Hz range. They normally include a high-pass filter to attenuate the undesired low-frequency noise below 80 Hz. Noise can also be found in the high-frequencies, mainly attributed to non-linear distortions of the acoustic system and to its vibration effects, especially in the case of small devices producing high sound pressure levels (e.g. speakerphones). Speech processing techniques may be applied in the device pursuing the improvement of the audio quality, somewhat altering the speech signal. Typical techniques are voice activity detection (VAD), noise reduction, acoustic echo cancellation (AEC), or comfort noise generation (CNG). Non-stationary noises such as wind noise or cafeteria noise are particularly challenging for the existing speech enhancement methods [63].

The relevant aspects of terminals affecting the transmitted signal are referenced in the ETSI standard method for end-to-end (mouth to ear) speech quality testing [65]. However, devices are not consistent between brands from the design and technology

point-of-view. While a large number of user interface components are standardized, particular devices vary in the applied speech enhancement approaches, with often unknown details. The influence of handsets and headphones in receiving direction in conjunction with that of different bandwidths has been found to be significant regarding signal quality [210].

2.2 Channel Quality Evaluation

Together with the expansion of the telecommunication infrastructure and the sophistication of speech processing algorithms there is the need for speech quality evaluation, often during the design phase of communication channels. The purpose of the quality assessment is to detect both how the communication is perceived by its users and how the needs and expectations of the users evolve. Speech quality is one attribute of the speech signal which consists of dimensions such as intelligibility, comprehensibility, naturalness, clarity, pleasantness, brightness, etc.

The speech quality can be assessed by performing listening tests, where a group of listeners listen to processed speech and rate their quality employing a pre-defined scale [124, 125], or by using instrumental quality measures, which quantify the difference in quality between the original and the processed signals. Because the auditory assessments are generally costly and time consuming, research has focused on the design of instrumental measures which can reliably predict the subjective rating scores. Instrumental measuring methods could then replace the subjective quality tests in the design of the deployed channel. For instance, the measure could be applied iteratively for the optimisation of the system parameters before it is offered to the market or be used to monitor and optimise a coding procedure dynamically [45]. Compared to human subjective tests, these are quicker, cheaper, more consistent, and not subject to human errors. However, human auditory assessments are sometimes preferable as they reflect more reliably the subjectivity of the listeners.

2.2.1 Subjective Speech Quality Assessment

Formal auditory tests for quality assessments are relevant for obtaining reliable quality ratings of the transmission link or speech processing system under deployment, e.g. a new codec or a new device. The listening tests follow the test methods of the ITU-T P.800 series of Recommendations [124, 125]. A panel of listeners judge the speech quality on an Absolute Category Rating scale ranging from 1 ("bad") to 5 ("excellent") or with reference to other speech samples (Comparison Category Rating or Degradation Category Rating). In Absolute Category Rating tests, the listener ratings are averaged to obtain the subjective Mean Opinion Score (MOS), which represents the resulting quality of each speech sample.

The effects of bandwidth on the perceived signal quality was examined in [265] in the absence of coding distortions. The author designed an auditory test based on paired comparisons to measure the subjective speech quality of 19 band-pass filters spanning the range of bandwidths between NB and WB. With NB being the reference in each pair of stimuli presented and with the second segment processed by one of the band-pass filters, the listeners rated the quality on a 7-points scale from -3 to $+3$. It was found that the WB bandwidth offered the best quality improvement over NB among the band-pass filters, that frequency ranges with the lower limit below 300 Hz offered better quality, and that the band 300–7000 Hz, that is, extending only the upper NB limit, offered quality comparable to NB, which reveals the critical importance of the low frequencies.

Another formal listening test to quantify the difference in perceived quality of different band-pass filters [211] showed that the degradation introduced by a band-pass filter decreased almost linearly with the extension of the bandwidth, corresponding the quality improvement of WB over NB to about 30 %. A quantitative model derived from the results of the subjective tests revealed that WB can already offer quality advantage over NB with a codec operating at a bitrate as low as 20 kbit/s. An extension of this study is presented in [267] towards SWB, which is shown to improve 39 % over WB and 79 % over NB for clean channels (when no codec is applied).

NB, WB, SWB, and Full-band (FB, 20–20,000 Hz) conditions along with different codecs were also examined in a subjective test in [214], where listeners gave their ratings on a modified MOS scale (9 "excellent" and 1 "very bad"). A significant improvement was shown as the signal bandwidth increased from NB to WB and from WB to SWB, although no significant benefits were offered by FB over SWB. Interestingly, listeners judged mono samples in SWB to offer better quality than WB stereo samples. The differences in quality between different wireless codecs and different ITU-T codecs were also shown, demonstrating the quality improvement achieved with the AMR codecs over other codecs of the same bandwidth. A complete analysis showing the quality improvements gained with the transition from WB to SWB is given in [271], focusing on quality evaluations of a variety of SWB codecs.

Three common perceptual quality dimensions relevant for NB and in WB speech were identified through listening tests in [268]: *Discontinuity*, *Noisiness*, and *Coloration*, assumed to cover the whole NB speech quality space, whereas a WB-specific dimension is added to the last three to cover the WB speech quality space. These perceptual dimensions are the base to estimate the quality degradations introduced by speech transmissions [48].

The influence of packet loss in VoIP communications on the perceived signal quality depends on factors such as channel bandwidth, codec applied, loss pattern, burst loss size, and location of loss within the speech [113, 210]. These factors are investigated to measure the degree of user satisfaction and also to assist the design of efficient speech recovery systems. The human quality ratings seem to decrease rapidly from a random packet loss rate of 5 % [210].

The electro-acoustic user interfaces at both ends of the transmission incorporate transducers (i.e. microphones and loudspeakers) presenting particular filtering characteristics that alter the speech quality. Besides, their geometry and the gap

between handset and listener's ear provoke a signal loss in the frequencies below 7 kHz [210]. It was found in [190], applying band-pass filters with different shapes, that the quality was significantly degraded when the lower limit of the frequency range increased from 123 to 208 Hz or when the upper limit fell below 10,869 Hz. Naturalness decreased progressively as the upper limit was lowered from 10,869 to 3,547 Hz, approximately the upper limit of NB transmissions. Spectral ripples with a depth of 10 dB, common in medium quality headphones, degraded naturalness more severely when they extended over a wide frequency range (87–6,981 Hz) than over frequency sub-ranges. The listening tests performed in [210] showed that MOS scores were higher employing a Hi-Fi phone for listening than employing a diotic headphone in NB. This outcome was reversed in WB, indicating the influence of the listeners' expectation towards different devices.

2.2.2 Instrumental Speech Quality Measures

Instrumental speech quality assessment is required for the design and management of networks and terminals when subjective tests are excessively time-consuming and expensive to run, and should predict the listener ratings as accurately as possible. The different models to estimate speech quality measures can be classified into signal-based models, which employ transmitted signals acquired at the receiver user interface, and parametric models, which can operate from technical specifications of the network during its design phase. Some of these specifications are the frequency-weighted insertion loss, delay, noise power, packet loss probability, and employed codec [187]. There also exist hybrid approaches to instrumental quality estimations employing network parameters and the transmitted signal. The signal-based models are further divided into "intrusive" or "non-intrusive", depending on whether both original and degraded speech signals are needed to compute the quality estimation or only the degraded version. The non-intrusive speech quality estimation methods may include a speech production model to identify the speech components to be separated from the artificial channel distortions. Only the instrumental measures most recent and relevant to transmission channels are presented in this review.

Perceptual Evaluation of Speech Quality (PESQ) is an intrusive signal-based model. Its NB version is specified in ITU-T Rec. P.862 [126] and its WB extension in ITU-T Rec. P.862.2 [127], which involves a different input filter and a different mapping function to MOS. The works in [113, 158, 245, 244] applied the PESQ model to analyse the NB and the WB VoIP quality indicating the superiority of WB communications. The studies [113] and [245] show the quality degradations introduced by different packet loss rates.

The Perceptual Objective Listening Quality Assessment (POLQA) model, described in ITU-T Rec. P.863 [129], is another intrusive model, successor of PESQ. It can operate in NB mode and in SWB mode, the latter covering a bandwidth wider than that considered in WB-PESQ and taking into account the electro-acoustic characteristics of the acoustic interfaces. It was shown recently in [109] that the correlation between POLQA and the subjective MOS was higher than for PESQ for WB data.

Apart from obtaining estimations of perceived quality, the detection of causes for the degradation of the transmission is also relevant for many scenarios. The three perceptual quality dimensions mentioned before for NB and for WB; *Discontinuity*, *Noisiness*, and *Coloration* [268] were the basis to develop multidimensional instrumental reference-based quality models such as the Diagnostic Instrumental Assessment of Listening quality (DIAL) model, also intrusive, which also includes the *Loudness* dimension in the case of non-optimal listening level [47, 48].

Regarding parametric models, the most widely used is the E-model, described in ITU-T Rec. G.107 for NB [117] and G.107.1 for WB [118]. It works on the basis of parameters describing each element of the transmission channel estimating the relative voice quality for a reference connection. Its primary output is the transmission rating R, or R-factor, ranging from 0 to 100 for NB, which can be transformed to an overall quality MOS. The R-factor was extended to WB in [189], where the range 0–129 was proposed, and to SWB in [269], reaching the range 0–179. The extrapolation of the E-model transmission rating scale was based on impairment factors derived from listening tests, and served also to estimate the quality of a variety of codecs at different bitrates for the three bandwidths. The E-model was recommended by ITU-T for network planning purposes, although it can also be employed for quality monitoring [45]. A parametric formula derived from the E-model, presented in [210], can be used to quantify the packet loss impairment.

2.2.3 Relations Between Quality and Other Attributes of the Speech Signal

The studies presented so far have demonstrated that extended signal bandwidths offer better quality. In addition, WB communications are expected to enable a better identification of phonemes, intelligibility, over NB [133, 222]. However, no formal listening tests quantifying this possible benefit are known to the author.

It has been shown that the PESQ model can predict human speech intelligibility [17], although modifications need to be introduced in the model in order to obtain more reliable estimations. The analysis presented in [169] did not find strong correlations between instrumental measurements of speech quality and human speech intelligibility. The authors degraded the voice signals applying various types of background noise with different SNRs and NB codecs, and reported even weaker correlations when speech enhancement schemes were applied, suggesting that these hampered intelligibility. An instrumental measure of speech intelligibility, called Coherence Speech Intelligibility Index (cSII), was shown to be less valid than other quality measurements to predict intelligibility for speech degraded by additive noise and by non-linear distortions [256]. The development of more reliable intelligibility models is still under investigation, many of them being targeted to predicting the intelligibility of synthesised speech for Text-to-Speech applications. Comparisons of quality and intelligibility over different channel bandwidths have been overlooked so far.

Some investigations have addressed the relationship between signal quality and automatic speech recognition. An attempt to develop an alternative model for instrumental quality assessment by employing an automatic speech recogniser was made in [43] and extended in [258] to estimate the quality of VoIP communications. The analysis in [112] shows a good correlation between MOS and automatic word recognition accuracy for the AMR-NB codec at different bitrates and the study in [136] for transmissions under packet loss degradations. Contrariwise, PESQ was found to be a good estimator of the performance of automatic speech recognition systems [255], as was the E-model, with the adjustment proposed in [221]. This adjustment was necessary because the model was originally optimised to predict quality in communications between humans. The study in [213] addressed the relationship between MOS and automatic speech recognition systems in GSM and VoIP networks, comparing NB and WB codecs. It was reported that MOS values were more affected by low bitrate coding than the automatic speech recognition performance.

Speaker recognition, as distinct from speech recognition, has also been related to signal quality and intelligibility in previous investigations. The relatively old work in [253], aimed at comparing the human speaker recognition ability and human speech intelligibility for real radio communication links, did not find evidence that channel impairments affected intelligibility and speaker recognition to the same extent. The work in [243] presents a comparison of automatic speaker recognition performance and MOS values over a variety of NB codecs, finding only a weak correlation between both speech attributes. Differently, the PESQ measure was found to correlate well with automatic speaker recognition under different distortions introduced by the Voice over Wireless Local Area Network (VoWLAN), GSM and PSTN networks, which proved useful for the prediction of speaker recognition performance in telephony [25]. It was suggested in [177] that the measurement of spectral distortion caused by NB coding could be used to predict automatic speaker recognition scores, although this prediction was not directly addressed.

Subjective speech quality assessments [189, 263, 269] and instrumental measurements [113, 158, 245, 244] have shown that signal quality improves when the signal is transmitted through a channel of extended bandwidth. They have also evaluated the degradations due to the codec, packet loss, and user interface, which are the main artefacts of PSTN and VoIP networks. In this book it is examined the correspondence between speech offering certain estimated quality and the speaker recognisability rate obtained with that speech. It is described how the latter can be predicted with certain reliability from different signal quality measurements.

2.3 Human Speaker Recognition

Speaker recognition is intuitively performed by humans in everyday situations when they associate a voice they hear to a voice heard before and by some means encoded in memory. For example, a known person can be recognised after listening to him/her speaking from another room, to his/her interview over the radio, or, more relevant to this research work, when listening to his/her voice through a telephone connection.

The main sources of error affecting the capacity of listeners to recognise voices are human-related and technical (dependent on environmental conditions). Human-related factors can be attributable to the speaker and to the listener. A correct recognition can be hampered by the speaker's physical condition (health or unusual speaker emotions such as anger), manner of speaking (pronunciation pattern, choice of vocabulary), and cooperativeness (intentional voice disguise). The listener's age, hearing ability, the length of the delay between initial exposure to a voice and the identification task, the human memory, the familiarity with the voice, and the length and the content of the sample heard may also have an influence on speaker recognisability. Technical error sources refer to the distortion introduced by transmission channels (for instance, type of handset used, channel band-pass, codec, line and switching equipment) and by the background noise. The mentioned factors cause non-desirable speaker variabilities, that is, the voice of the same speaker may sound different when human or technical conditions vary from one sample to another. This sub-chapter reviews auditory tests that identify these factors and measure their influence, with applications in forensic speaker recognition, cochlear implants, and telephony, among others.

Controversy in a legal case was generated in 1935 where an earwitness recognised the subject's voice some years after hearing a perpetrator. This issue stimulated the commencement of formal research on the validity of human voice identification and influential factors, as early as 1937 [178], finding decreases in unfamiliar-speaker recognition accuracy after varying intervals of time. The research on how speaker identity is encoded by listeners, originally relevant to forensic investigations, was then extended to other fields such as linguistics, psychology, speech science, and audiology (particularly hearing impairment and cochlear implants). It has evolved at a rapid pace along with the developments in speech processing mechanisms and telephony. For instance, analyses of forensic speaker recognition need to consider new artefacts affecting voice recordings [225], and cochlear implants can be based on new and more sophisticated voice processing schemes [264]. One of the later applications of human speaker recognition studies is to assist the design of efficient automatic speaker recognition systems [82].

2.3.1 Speech Characteristics Enabling Human Speaker Recognition

Depending on the research objective, listening tests may examine the human speaker identification (SI) or the human speaker verification (SV) ability. SI is the task where listeners detect a speaker among several given possibilities and are, thus, already familiar with the voices. This is a natural situation in a phone call scenario, relevant to the present work. In contrast, SV is performed when listeners are asked to compare two generally unknown voices or to rate their similarity. This task is of interest for forensic studies [221].

Features or cues employed by listeners to recognise speakers were examined in numerous investigations in the last decades to clarify the mechanisms of human perception and recognition of voices. The term "voice quality" refers to the auditory colouring of a particular voice resulting from laryngeal and supralaryngeal activity, e.g. nasal, whispery, or breathy voice. This includes, for instance, pitch, loudness, breathiness, laryngealisation, and phonation types, which are commonly altered or removed in auditory tests to demonstrate their effectiveness for the performance of listeners recognising voices [15, 26, 27, 159, 161, 163, 215, 262]. The most useful parameter, emerging across all studies, is the fundamental frequency (F0).

It has been asserted that the critical parameters or acoustic cues for correct speaker recognition depend on the particular voice heard, which may be more or less distinctive. One of the earliest investigations to draw this conclusion conducted listening tests where voices of famous people and with different lengths were played forward and backwards [161]. While up to the date of this study it had been believed that F0 and F0 contour were primary cues to speaker recognition, the authors indicated pitch characteristics were only effective for the recognition of some voices. An overall decrease of 12 % in correct identifications from forward to backward stimuli presentation was reported, which suggested that speaker recognition could already be successful from acoustic parameters such as pitch and pitch range, speech rate, voice quality and vowel quality; and without the presence of articulatory and phonetic patterns or temporal structure. The influence of F0 height, F0 contour, and speech rhythm was later analysed in [262], also confirming that the perceptual importance of the pitch parameters depends on the target voice to be identified. The author found that voices with average pitch were less sensitive to variations of F0 height than voices with low or high pitch, although this is presumably highly dependent on the familiarity of the listener with the voice.

A speaker verification experiment where listeners rated the similarity between two voices on a 7-point scale was presented in [159]. It revealed that, together with F0, listeners may also utilise other salient acoustic parameters, such as variations of F0 and of loudness, when these exhibit great variability. The extent of the deviations between listeners' ratings depended on the heterogeneity of the voices of the test. Employing only vowels as stimuli, the results in [15] suggested that the dispersion between the fourth and the fifth formants facilitated more accurate differentiation of male speakers while the first formant was of greater importance for female speakers, due to the lower energy of the higher formants. The authors also showed, consistently with the previous literature, that F0 was the principal parameter for correct speaker recognition from vowels. Shifting the third and the fourth vowel formants towards lower frequencies had a greater effect on the speaker identification performance than increasing those frequencies and that varying the first and the second formants [163]. Speaker glottalisation [203, 273], defined in [27] as *the rate and type of intermittent irregular vocal fold vibration* was also found to be an important cue for listeners identifying familiar and unfamiliar speakers [26, 27].

Listeners do not only rely on acoustic correlates of voice quality to identify familiar and unfamiliar voices, but also on phoneme articulations. These appeared to be characteristic of the speakers and are crucial as well for the task of word recognition [215]. The effects of the phonological content of stimuli on talker recognition have been examined extensively. Early studies indicated that the speaker identification performance improved with stimuli of longer durations due to an increased number of different phonemes being uttered by the talker [28, 204]. The degree to which different speech sounds convey speaker-specific characteristics was later investigated in [61, 231, 270] employing automatic speaker recognition. The authors agreed that vowels and nasals provide the best discrimination between speakers. It has also been asserted that fricatives contribute to the speaker recognition performance to a lesser extent, and that stop sounds are the least useful phoneme category for that purpose [61]. Also, there is evidence that each talker may produce different speaker-discriminative sounds [176]. Regarding human talker identification, the importance of vowel sounds is commonly acknowledged. Front stressed vowels [250] and nasalised vowels [4] have been found to be particularly useful. The works in [5–7] have evinced that nasal sounds facilitate higher human speaker recognition than other consonants. This is attributable to the fact that the resonance cavities shaping nasal sounds differ considerably among speakers [251].

It is assumed that human speaker recognition and human speech intelligibility are closely interrelated, as both are performed from linguistic cues. An in-depth review of this is given in [50] from a psychological point of view. The human capability to recognise speech has often been evaluated with rhyme tests, Semantically Unpredictable Sentences tests, or Cluster-Identification tests, which measure the comprehensibility of words or monosyllables previously altered depending on the tests objectives. Humans are able to recognise short speech segments with little or no high-level grammatical information [168]. Background noise is considered one of the main factors affecting speech intelligibility of logatomes (nonsense syllables in the form vowel-consonant-vowel (VCV) or CVC) and from CV syllables [104, 180, 181, 183, 201]. In quiet conditions, [201] (in English language) showed that the most confused consonants were the fricatives /ð/-/θ/, /ð/-/v/, /ʒ/-/ʃ/, /θ/-/f/ and, to a lesser extent, the stop sounds /p/-/b/.

It must be underlined that the evaluation of the cues mentioned above only applies to speaker recognition performed by humans, while automatic systems do not necessarily use the same cues to identify or to verify speakers. Outcomes from listening tests have partially inspired the design of automatic speaker recognition algorithms and can also complement the machine performance, as shown by the NIST HASR challenges [240]. However, automatic systems employ different procedures and speech training material than humans, who rely on their exposure to the speaker's voice, memory, and life-long experience distinguishing among speakers, and it is well known that automatic systems perform generally better than human listeners in the speaker verification task, at least in the absence of background noise [3].

2.3.2 Effects of Communication Channels on Human Speaker Recognition

Voice transmissions through communication channels presenting different characteristics add additional variations to the voices and may modify the speakers' salient features reviewed before, hampering their correct recognition.

Although no channel transmission was involved, early studies have revealed, employing different band-pass filters, that frequencies above 1 kHz carry more speaker-specific information than the lower frequencies [46, 200], while speaker recognition rates were still well above chance level when words were high-pass filtered at 5 kHz or low-pass filtered at 100 Hz [204]. Of the frequency ranges examined, it was shown that the band of approximately 1–2.4 kHz was the most beneficial for human speaker recognition [200, 204]. The studies in [237, 238, 261] examined the listener's performance when voices were transmitted through NB Linear Predictive Coding (LPC) voice processors. The work in [237] reported a decrease of the listeners' accuracy identifying 24 familiar speakers from 88 to 69 % when the voices were transmitted, which was considered acceptable for that voice communication system. The reference [261] showed that a LPC voice processor with band-pass 200–3,500 Hz offered significantly better performance than another with band-pass 100–3,000 Hz, corroborating the effectiveness of high-frequency components over the lower frequencies. In particular, the higher frequencies carry information of voice quality and specific phonation types, which, as mentioned before, are distinctive characteristics of each person.

Speaker verification listening tests were found appropriate and adopted for evaluating a new American Department of Defence standard codec in terms of speaker recognisability [234, 235]. It was concluded in [235] that, in contrast to other approaches for forensic speaker recognition, the comparison between two stimuli should be in the form unprocessed-processed and processed-processed in order to assess the codec's capability to both preserve speaker characteristics and to permit speaker discrimination. A more recent evaluation of human speaker recognition under transmissions through assorted speech codecs was conducted in [40]. The authors employed the Improved Multiband Excitation Codec (IMEC) and the Mixed Excitation Linear Prediction (MELP) codec (both of them NB) at low bitrate, and other distortions such as noise and packet loss. Their outcomes suggested that speaker identification was significantly less affected by channel degradations than speech intelligibility. However, the influence of common codecs employed in landline or in VoIP telephony, such as the G.711 or the G.722 has not yet been assessed for human speaker recognition.

One single work has been found that analysed the differences between the standard bandwidths NB and WB for human speaker recognition, leaving aside the effects of speech codecs [62]. It was demonstrated that the speaker verification accuracy decreased more rapidly when the cut-off frequencies of low-pass filters fell inside the WB range than when they did beyond. Besides, it was argued that WB channels allowed humans to recognise speakers with similar accuracy to full-band speech.

The current tendency of human speaker recognition studies is focused towards the ability of humans to compare pairs of voices in a forensic context. To date, the National Institute of Standards and Technology (NIST) has organised Human Assisted Speaker Recognition (HASR) challenges in the years 2010 and 2012, as part of a series of Speaker Recognition Evaluations (SREs). The intention has been to investigate whether decisions made by automatic systems on same/different speaker can benefit from human judgements. However, no attention is paid to the possible effects of different transmission configurations. Although only little interrelation between automatic and human decisions was found [95], human-machine fusion seems to be promising to strengthen both performances [240]. Crowdsourcing via Mechanical Turk has shown to be an effective approach for obtaining large-scale human ratings by listening, well-matched with forensic experts' decisions [242].

The SRE challenges have been proposed to the speaker recognition community since 1996 with the aim of establishing biometric standards, which would permit to compare the performance of different automatic speaker recognition systems by defining common datasets and methods to assess the performance. The systematic benchmark tests facilitate a collaborative work between the researchers, who achieved enormous progress in the last decades. The NIST SREs will be addressed in more detail in the review of automatic speaker recognition (in Sect. 2.4.4).

2.3.3 Literature on Human Speaker Recognition and This Book

Several investigations have asserted that the human voice has important speaker-specific content beyond the cut-off frequencies of NB channels. However, the performance of listeners recognising voices transmitted through WB channels has not yet been evaluated. This book examines the comparison between performances over different bandwidths (NB, WB, SWB) and the relationships between human speaker recognisability, speech intelligibility, and quality of audio signals for different channel degradations.

As reviewed, aspects of the listening test set-up such as the number of speakers to be identified, the listener familiarity with the voices, and the stimuli content and duration may have a strong influence on the listeners' performance recognising speakers. These factors are considered carefully in the design of the listening tests developed for this book, with the intention of obtaining appropriate results and listeners' accuracies in an adequate range, far from chance level and from saturation, which will enable the comparison between different transmission conditions.

2.4 Automatic Speaker Recognition

The automatic detection of people's identity from their voices, without requiring the intervention of humans, has attracted the attention of researchers and engineers in the last decades. Humans are able to genuinely discriminate between voices without

difficulties if these are previously known and not severely distorted. In contrast, automatic speaker recognition systems require a meticulous statistical classifier design and a careful selection of training data and speech enhancement techniques, depending on the environmental condition and application requirements. Automatic speaker verification (ASV) consists of the determination of whether two given utterances originate from the same speaker or not. It generally performs the validation of an individual's identity by first learning the voice of the target speaker (enrolment phase) and then accepting or rejecting an identity claimed by either a legitimate speaker or an impostor (verification phase). While humans commonly identify a given utterance among a learned set of voices (speaker identification), speaker verification is the general task performed by automatic systems.

The interest in automatically recognising talkers began in the 1960s [208], one decade later than for automatic speech recognition. One of the main applications of ASV since then is biometrics, i.e. secure access control by voice. In addition, comparing pairs of voices is extremely useful in a forensic context [225], where the system's judgements are often more accurate than the true/false decisions made by humans [3]. The automatic recognition of identities is also valid for services and systems customisation, speech data management, and surveillance. In the case of multispeaker recordings, the detection of who speaks when given an audio stream (speaker diarisation), benefits from the ASV technology [260]. Besides speaker recognition, there exist other speech processing technologies that are concerned with the extraction of information, other than speaker identity, from the speech signal. Examples of these technologies are: speech recognition, gender identification, detection of the talker's emotion and personality, and language, dialect, and accent recognition.

One of the major challenges of ASV systems is the difference in nature of the enrolment and test material, which causes a decrease in performance. This mismatch may not only come from background noise and channel degradations, but other factors such as the physical and emotional state of the talkers also contribute to within-speaker variations. Currently, the automatic systems still deal with the arduousness of separating the environmental characteristics (technical- and speaker-related errors) from those of the speaker. For instance, a legitimate user employing a telephone connection different than that used for enrolment may have lower chance of being accepted. It can also occur that an impostor whose voice is transmitted through the enrolment channel of rightful clients is mistakenly authenticated. The mismatch is particularly severe in the typical case where the segments for enrolment are recorded directly from a microphone and the voice for verification transmitted through telephone. Besides mismatch, performance is also affected by the natures of the enrolment and test data in terms of speech duration and recording characteristics (which may vary among the sessions of each speaker in the enrolment set).

Although ASV is not entirely secure and is sometimes combined with other biometric modalities such as face recognition when high performance is required [44], the low cost and the non-intrusive nature of the input device facilitate its widespread use in commercial applications. Speech-based person authentication is relevant in services such as voice dialling, phone banking or mobile-phone purchases. ASV can be classified into text-dependent, where speakers are asked to utter a given text

prompt, or text independent, without constraint on the speech content. This book focuses on the latter approach, which provides more flexibility to the system yet is more challenging because has less control over the user input. Speaker recognition employed for forensic investigations faces other problems such as uncontrollable recording conditions, emotional speech, and uncooperative speakers. Indeed, voice disguise and voice mimicry can mislead automatic systems [70, 162]. Efforts have been made to combat spoofing attacks by detecting playbacks of recorded speech, voice transformation and synthesised speech [2, 266].

2.4.1 Automatic Speaker Recognition Principles and Main Systems

The basic operation principle of automatic speaker recognition is to extract speaker-specific features from the speech signal and to apply some sort of modelling technique to effectively represent this information. The decision of whether two voices correspond to the same or to different speakers is made by comparing the similarities between already enrolled speaker models and a given utterance at verification time. Speaker verification systems in the literature propose different techniques to compute the likelihood of the given utterance being spoken by the hypothesised speaker and the likelihood of that utterance spoken by another speaker. A certain threshold θ, depending on the application requirements, is compared to the quotient between these likelihoods to decide whether the hypothesised speaker is authenticated or not. Assuming that the models are well estimated, the computation of the likelihood ratio is the optimal way to make such speaker verification decisions.

H_0 the utterance Y corresponds to the hypothesised speaker S
H_1 the utterance Y *does not* correspond to the hypothesised speaker S

$$\frac{p(Y|H_0)}{p(Y|H_1)} \begin{cases} \geq \theta \text{ accept } H_0 \\ < \theta \text{ reject } H_0 \end{cases} \tag{2.1}$$

It is desirable that the features extracted from the speech for later modelling be easily measurable, occur frequently in speech, and present relevant speaker information. They should present large between-speaker variability and small within-speaker variability, and not be affected by speaking manner or by noise and transmission characteristics [157, 254, 270]. Depending on the type of information offered, they can be classified into short-term, voice source, spectro-temporal, prosodic and high-level features. While the low-level acoustic information offered by the short-term spectral features is employed by most systems, the combination with higher-level information, such as prosody, speech rate, word usage and other suprasegmental features can significantly improve the system's performance [1, 217]. However, these prosodic features have not been very successful by themselves for automatic speaker recognition compared to low-level features [1]. The most widely employed sets of

cepstral features are Mel-Frequency Cepstral Coefficients (MFCCs) [51] and Perceptual Linear Prediction coefficients (PLP) [107]. They are commonly extracted by first partitioning the signal into speech frames applying a window of 25 ms with increments of 10 ms. The non-speech frames are then removed by a Voice Activity Detector and feature warping [197] can be applied to compensate for channel variability. 10 to 20 coefficients are typically extracted and first and second derivatives (delta and delta-delta cepstra) can be appended to form the feature vector. Common noise-robust features applied to speaker recognition in noisy scenarios are Mean Hilbert Envelope Coefficients (MHEC) [228], Medium Duration Modulation Cepstrum (MDMC) [186], and Power Normalized Cepstral Coefficients (PNCC) [154].

Over the decades of automatic speaker recognition research the preferred modelling techniques have varied from those applied for text-dependent systems to those effective for text-independent applications. Approaches widely implemented in the 60s, 70s, and 80s were spectral template matching, Dynamic Time-Warping (DTW), and Vector Quantisation (VQ). Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) were very widely used in the 90s for text-dependent and for text-independent approaches, respectively. The HMM states represent the allowed speaker utterances, whereas the GMM components, commonly 512, 1024, or 2048 represent speaker spectral individualities [219]. The GMM-Universal Background Model (UBM) was introduced in 2000 [218] and is still the base of many current speaker recognition investigations.

A GMM is a parametric probability density function that models the distribution of feature vector sequences. It is given by a weighted sum of M component Gaussian densities:

$$p(\mathbf{X}|\lambda) = \sum_{i=1}^{M} w_i \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2.2}$$

where $\mathbf{X}$ is a feature vector, $w_i, i = 1, \ldots, M$, are the mixture weights, and $\mathcal{N}(\mathbf{X}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, \ldots, M$, are the unimodal Gaussian densities of mean vector $\boldsymbol{\mu}_\mathbf{i}$ and covariance matrix $\boldsymbol{\Sigma}_\mathbf{i}$. The mixture weights satisfy $\sum_{i=1}^{M} w_i = 1$. The speaker GMM, whose parameters are estimated from a collection of training feature vectors using the iterative algorithm, is denoted as: $\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}, i = 1, \ldots, M$.

It was proposed in [218] that a UBM, which is a speaker-independent GMM that represents the population of alternative speakers, can be used to derive the speaker-dependent models from the enrolment data of target speakers. The UBM should be trained with speech from a large set of non-client speakers to represent general speech characteristics. These data should reflect the expected alternative speech to be encountered at enrolment and verification time, i.e. ideally, when sufficient prior information is available, these signals should present the same type of transmission channel distortion and speaker characteristics such as gender, speaking style, and language.

The GMM client models λ_S are derived by adapting the UBM parameters λ_{UBM} to those of a target speaker by means of the Maximum *a Posteriori* (MAP) adaptation, also described in [218]. The log-likelihood ratio (LLR) is then computed given the feature vectors $\mathbf{X}$ extracted from the test utterances as:

$$LLR(X) = \log p(\mathbf{X}|\lambda_S) - \log p(\mathbf{X}|\lambda_{UBM}) \tag{2.3}$$

where the speaker model λ_S and the UBM λ_{UBM} represent the distribution of acoustic features for the target speaker S and for the general population, respectively. The LLR is compared to a given threshold to accept or reject the claimant, as similarly indicated in the expression 2.1.

Later, new techniques emerged which addressed the reduction of dimensionality and of intra-speaker variability, showing improved performance under session variability (e.g. different recording conditions). The speaker utterances were represented by supervectors, which consist of the concatenation of speaker-dependent GMM mean vectors of given training samples. This made it possible to work directly with vector-matrix manipulations. For instance, the GMM supervectors were used to derive a Support Vector Machine (SVM) kernel for speaker classification [37, 131]. The Nuisance Attribute Projection (NAP) technique could then be further applied to diminish the problem of session variability [38]. One drawback of supervectors was their high dimensionality, typically $\approx 60 \times 2048$.

The Joint Factor Analysis (JFA), proposed for the GMM frameworks in [146, 148] allows to create session-compensated speaker models by separating the speaker characteristics from the so-called nuisances or channel characteristics, which were modelled in the speaker space and in the channel space, respectively [150]. According to this approach, a supervector $\mathbf{M}$ can be decomposed into a sum of speaker- and channel- or session-dependent contributions:

$$\mathbf{M} = \mathbf{s} + \mathbf{c} \tag{2.4}$$

where $\mathbf{s}$ and $\mathbf{c}$ are referred to as the speaker-dependent and channel-dependent supervectors, respectively, and are described as:

$$\mathbf{s} = \mathbf{m} + \mathbf{Vy} + \mathbf{Dz} \tag{2.5}$$

$$\mathbf{c} = \mathbf{Ux} \tag{2.6}$$

where $\mathbf{m}$ is the speaker- and channel-independent UBM mean supervector, $\mathbf{V}$ and $\mathbf{U}$ the eigenvoices and the eigenchannel matrices, defining the speaker and the channel space, respectively, and $\mathbf{D}$ a diagonal matrix. $\mathbf{y}$ and $\mathbf{x}$ are vectors with components referred to as speaker factors and channel factors, respectively, and are assumed to have standard normal prior distributions. The term $\mathbf{Dz}$ serves as a residual.

The JFA matrices can be trained by extracting the Baum-Welch statistics from the acoustic observations, or feature vectors, and then iterating the *maximum likelihood re-estimation* and *minimum divergence re-estimation* processes to estimate the hyperparameters $\mathbf{V}$, $\mathbf{U}$, and $\mathbf{D}$. The $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{z}$ factors are extracted from the computed matrices. Large amounts of audio data are generally required, recorded under diverse conditions foreseen to be contained within the evaluation data [148].

300 eigenvoice and 100 eigenchannel components are typically estimated. Exhaustive descriptions of the JFA model and its training and evaluation procedures can be found in [146, 151].

In the special case $\mathbf{y} = 0$, the speaker supervector $\mathbf{y}$ describes the MAP adaptation technique of the standard GMM-UBM approach [218]. Hence, the JFA model can be seen as an extension of that technique as it combines classical MAP, eigenvoice MAP and eigenchannel MAP to adequately model the additive speaker and channel effects. The GMM-UBM weakness in comparison with JFA is the adaptation of not only speaker-specific characteristics, but also the channel variability and other nuisance factors implicit in the background population speech of the UBM. This is done ideally in separate adaptations with the JFA approach. In addition, it allows a client model to be approximately represented by the speaker factors $\mathbf{y}$, of lower dimension than the supervectors, facilitating enrolment with limited data. SVM combined with JFA was shown to be an efficient modelling technique [54].

Despite the acceptable performance of JFA, it was found in [52] that some useful information to discriminate between voices was contained in the JFA's channel space, which led to the development of the *identity vectors* or, commonly, i-vectors [53]. The i-vectors were proposed as a new approach to front-end analysis for SVM classification, yet it was found that fast-scoring techniques such as cosine distance scoring increased efficiency while provided performance similar to SVM [52]. These fast-scoring techniques also improve the more complex and time-consuming evaluation of the JFA likelihood expressions [148].

The i-vector paradigm can be seen as a feature extractor inspired by the JFA. Instead of the JFA approach of separately modelling the between-speaker and the within-speaker variability in a high dimension space of supervectors, a low-dimensional subspace of the GMM supervector space was proposed in [53], termed the total-variability space $\mathbf{T}$. It represents both speaker and channel variability. The vectors modelled in the total-variability space are the i-vectors, which can be represented by:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\boldsymbol{w} \tag{2.7}$$

where $\mathbf{M}$ is the speaker- and channel-dependent supervector and $\mathbf{m}$ the UBM mean supervector as in the JFA equations, $\mathbf{T}$ the total-variability matrix defining the total-variability space, and $\boldsymbol{w}$ an independent normally-distributed random vector representing the total-variability factors or i-vector. Typically, the $\mathbf{T}$ matrix defines i-vectors of dimension $\approx$400.

The $\mathbf{T}$ matrix can be computed by applying the *maximum likelihood re-estimation* iteratively as the $\mathbf{V}$ matrix of the JFA is estimated, with the only difference that the utterances corresponding to the same speakers should be regarded as having been produced by different speakers to train $\mathbf{T}$. The i-vectors can be extracted given the $\mathbf{T}$ matrix following a similar procedure as extracting the speaker factors $\mathbf{y}$ from the $\mathbf{V}$ matrix in the JFA approach. A covariance matrix $\boldsymbol{\Sigma}$, which models the residual variability not captured by $\mathbf{T}$, is employed in order to perform the i-vector inference, as in Eq. 6 of [53]. This covariance matrix was estimated during UBM training.

For the computation of the system scores, the straightforward cosine distance scoring technique can be applied:

$$score(\boldsymbol{w}_{target}, \boldsymbol{w}_{test}) = \frac{\langle \boldsymbol{w}_{target}, \boldsymbol{w}_{test} \rangle}{\|\mathbf{w}_{target}\| \|\boldsymbol{w}_{test}\|} \tag{2.8}$$

where $\boldsymbol{w}_{target}$ and $\boldsymbol{w}_{test}$ are i-vectors extracted from enrolment and from test material, respectively. This technique offers, thus, an easy approach to compare between sequences of features with different duration.

Channel compensation approaches are required within the i-vector space before the scoring phase to diminish the channel variability effects. Both non-probabilistic and probabilistic pattern recognition approaches have been proposed to decompose the signal into a speaker-specific component and a channel variability component. The most successful ones are Linear Discriminant Analysis (LDA) [58], Within-Class Covariance Normalization (WCCN) [101], Nuisance Attribute Projection (NAP) [39], and Probabilistic Linear Discriminant Analysis (PLDA) [207]. LDA is capable of projecting the i-vector feature into a much lower dimensional space, and of maximising the variance between speakers while minimising the intra-speaker variance [143]. The WCCN compensation technique is often combined with LDA to reduce the within-speaker variance [53].

The PLDA, with the same operation principle as the JFA, has been adapted recently from face recognition [207] to speaker recognition. There exist three variants of the PLDA model: the standard [207], the simplified [147], and the two-covariance PLDA variants [31]. They have been examined in [247], which concluded that the simplest possible model appropriate for the intended application should be used. Heavy-Tailed distributions (HT-PLDA) were shown to outperform Gaussian priors (G-PLDA) [147], although it was found in [85] that G-PLDA with length normalisation offered similar performance to that of HT-PLDA, with the advantage of being more efficient. This technique assumes that the i-vector Φ_{ij}, corresponding to the jth utterance of the ith speaker, is generated according to:

$$\Phi_{ij} = \boldsymbol{\mu} + \mathbf{S}\mathbf{y}_i + \varepsilon_{ij} \tag{2.9}$$

where the speaker-specific part $\boldsymbol{\mu} + \mathbf{S}\mathbf{y}_i$ describes the between-speaker variability as in the JFA approach, $\mathbf{S}$ constituting the eigenvoices (Eq. 2.5), and ε_{ij} is a residual term normally distributed with zero mean and full covariance matrix $\boldsymbol{\Lambda}_f$. It is denoted by the sub-index f that it is a full precision matrix [247]. The G-PLDA model parameters $\{\boldsymbol{\mu}, \mathbf{S}, \boldsymbol{\Lambda}_f\}$ are estimated from development data using the Expectation-Maximisation (EM) algorithm.

The new deep learning techniques are currently showing their strength in machine learning areas. While neural nets were trained discriminatively in their first uses, it has now been shown that, by adding an initial pretraining which ignores the fundamental system goal, they can achieve significant gains. For instance, Deep Neural Networks (DNNs) offer improved performance over conventional methods when integrated in speech recognition systems [110]. Recent investigations in speaker

recognition are incorporating DNNs for speaker modelling and showing competitive results [89, 165].

For the assessment of the system's performance two types of errors are taken into account: false acceptance or false alarm (FA), when an identity claim made by an impostor is falsely accepted by the system, and false rejections (FR), when a valid identity is falsely rejected. The rates of false acceptances and false rejections (P_{FA} and P_{FR}) define the operating point of the system, which can be established by setting a threshold in the decision making process. The Detection Error Trade-off (DET) curve is the plot of P_{FA} as function of P_{FR} on a normal deviate scale [173]. When a threshold has been set the system is evaluated in terms of the detection cost function (DCF) [175], which weighs the two types of errors by their respective costs (C_{FR} and C_{FA}). With the *a priori* probability of a target speaker occurring in the test set (P_{target}), the DCF can be written as:

$$DCF = C_{FR} P_{FR} P_{target} + C_{FA} P_{FA} (1 - P_{target}) \tag{2.10}$$

The system's operating point can be set so that the DCF is minimal (minDCF), which is a typical measure of the system's performance. Other standard performance measures are the equal error rate (EER), which corresponds to the operating point where $P_{FA} = P_{FR}$, and the half total error rate (HTER), which is the average of P_{FA} and P_{FR} at a specific threshold setting. Determining the system's operating point is a trade-off between FA and FR. When high security is required in practical applications such as phone banking the cost of FA is high. Setting a threshold that permits a low P_{FA} implies that the system will accept higher P_{FR} and vice versa.

Due to different recording conditions, duration of utterances, and phonetic content, it is often difficult to set a decision threshold. This problem is alleviated by score calibration, which applies an affine transformation on the score distribution to compensate for different intra- and inter-speaker score variations. Score normalisation techniques can calibrate the scores to some degree and thus improve the verification performance. Common score normalisation techniques are Z-norm, H-norm and T-norm, depending on whether the score distributions are estimated during the training phase or during the test phase and on the variability for which to compensate. A detailed description of score normalisation is given in [9].

2.4.2 *Effects of Phonetic Content on Automatic Speaker Recognition*

The success of an automatic system authenticating speakers is highly dependent on the content of the talkers' utterances with which it is confronted. Since some phonemes carry more speaker-discriminative information than others, it is important that the utterances include sounds that ease the speaker detection, especially when the duration of the speech at enrolment or at verification time is limited [172]. It was

already reviewed in Sect. 2.3.1. That different phonemes also affect the performance of humans recognising speakers [28]. This section presents a more detailed review of the location of speaker-specific characteristics in the speech spectrum found to be relevant for automatic speaker recognition and on how this information is extracted and employed.

A number of studies have demonstrated that vowels and nasals provide the best discrimination between speakers [61, 172, 231, 270], while fricatives and stop sounds contribute to the speaker recognition performance to a lesser extent. These findings have been applied to speaker recognition approaches that take advantage of the most speaker-distinctive sounds, aided by a phoneme detector [8, 11, 99, 137]. Although vowel sounds have proven to be effective for characterising individual speakers and been widely used for speaker recognition and in forensic analyses [16, 224], there is a growing interest in also exploiting the discriminative properties of fricatives and nasals.

Fricative consonants differ among speakers owing to their articulatory and acoustic properties [91, 114, 145]. Also, due to the complex and relatively fixed nasal and paranasal cavities of talkers [251], nasal consonants display low within-speaker and high between-speaker variability [64, 223, 251]. Nasal congestion and laryngeal inflammation, however, may create severe spectral perturbations affecting speaker verification results [226]. The importance of fricatives for speaker recognition was first reported in [196]. Interestingly, a recent study has shown that fricatives and nasals can be more useful than vowels for speaker discrimination [233]. The authors examined the speaker discrimination ability of phonemes applying different bandwidth filters and computing the F-ratios, which account for the relation between the variance of features between speakers and the variance within a speaker. Relevant to the work in this book, fricatives and nasals exhibit spectral peaks at high frequencies, from 3 to 8 kHz depending on the particular phoneme [140], which are suppressed in NB channels. The NB bandwidth filter eliminates also the important nasal content below 300 Hz.

Due to the occurrence of phonetic events with different spectral characteristics, information about the talker individuality is not equally distributed among the spectral sub-bands, that is, certain sub-bands present more discriminative power than others. The most discriminative frequencies found in different studies are shown in Table 2.1, along with the databases employed, indicating whether the speech was clean or distorted and its bandwidth. Only investigations performing sub-band analyses are included in this table. Overall, these studies agree that the lower frequency region (below 1 kHz) and the higher frequencies (above 3 kHz) provide better recognition accuracy than the middle frequencies. For instance, sub-band analyses have shown that vowel formants convey speaker individuality [19], particularly the third and the fourth formants [163, 231], which are manifested at higher frequencies for female speech due to their shorter vocal tract compared to males [69, 278]. Nasals present discriminative power in low and mid-high frequencies [115, 164], and other consonants in the upper part of the frequency spectrum, above 6 kHz [115].

To detect which frequencies convey speaker information the spectral domain is often partitioned into frequency sub-bands and their effectiveness for speaker

Table 2.1 Type of data and speaker-discriminative frequencies determined by sub-band analysis

References	Dataset (distortion, frequency range)	Findings: most discriminative sub-bands
[19]	TIMIT (clean, 0–8 kHz)	Below 0.6 kHz and above 2 kHz
[10]	Local set of 20 males and 13 females (clean, 0–4 kHz)	Below 0.6 kHz and above 2 kHz
[274]	NTT-Voice Recognition (clean, 0–8 kHz)	0–2 kHz and 6–8 kHz
[20]	TIMIT (clean, 0–8 kHz) and NTIMIT (NB, 0.3–3.4 kHz)	Below 0.6 kHz and above 3 kHz
[195]	TIMIT, 5th dialect region (clean, 0–8 kHz)	Below 1 kHz and 3–4.5 kHz
[155]	TIMIT (clean, 0–4 kHz)	0.05–0.25 kHz for all phoneme classes
[156]	TIMIT, 7th dialect region and Helsinki corpora (μ-law, 0–5.5 kHz)	Below 0.2 kHz and 2.5–4 kHz (TIMIT) and 2–3 kHz (Helsinki)
[246]	BT Millar speech database (clean, 0.3–3.4 kHz)	1–2.5 kHz and 2.5–4 kHz
[170, 171]	NTT-Voice Recognition (clean, 0–8 kHz)	0.05–0.3 kHz, 4–5.5 kHz, and 6.5–7.8 kHz
[164]	NIST SRE 2008 (μ-law, 0.3–3.4 kHz)	Around 0.3 kHz and above 2 kHz
[230]	Accent of British English (clean, 0–11.025 kHz)	Below 0.77 kHz and 3.4–11.025 kHz
[115]	RyongNam2006 (clean, 0–11.025 kHz)	Below 0.3 kHz, 4–5.5 kHz, and above 9 kHz

recognition analysed in different manners. For instance, in the study presented in [19] the authors applied a speaker recogniser to each sub-band separately and then combined their outputs to compute the global decision for text-dependent speaker identification. Some years later, they proposed an on-line feature selection procedure based on their analysis of the most discriminative frequency sub-bands [20]. In [246], the cepstral parameters from different sub-band systems were recombined with sub-band weighting. Optimum band splitting and recombination strategies were addressed in [274]. The authors of [10] employed linear and mel scale filters to analyse the sub-band discrimination power and developed a new frequency warping function (between linear and mel) that provided optimal speaker identification results employing a relatively small speaker dataset (20 males and 13 females).

Other investigations were concerned with the design of a custom filterbank as an alternative to the conventional mel-scaled filterbank to extract features that emphasise speaker-specific properties. In [195] the sub-band weights were determined using F-ratios and vector ranking criteria. This work was extended in [155] by adapting the weights of each sub-band depending on the phone detected in the input speech frame, that is, his proposed filterbank emphasised the discriminative power of particular phonemes. In [170] and in [171], the authors designed sub-band filters with non-uniform bandwidth which was inverse proportional to the F-ratio calculated on each frequency sub-band, whereas the filterbank developed in [115] was based on an

F-ratio study considering different phoneme classes. All of these studies showed that the features extracted with custom filterbanks outperformed the MFCC, which evidences that the latter might not be optimal for the task of speaker recognition. The work in [171] was extended in [164] for NB telephone speech, demonstrating the superiority of Linear Frequency Cepstral Coefficients (LFCCs) over MFCCs for the nasal and non-nasal consonant regions. Also [278] showed the advantages of LFCCs over MFCCs for NB speech, and that the benefits were accentuated for female speakers. Indeed, the superior resolution of the linear-spaced filters in the higher frequencies, where important speaker individuality is present according to these analyses, can capture more spectral detail and lead to better speaker recognition results compared to the mel-spaced filters.

The mel scale is based on human auditory characteristics and the MFCCs were originally developed for speech recognition and for signals band-limited to 5 kHz [51]. Hence, although this feature set is used extensively for the speaker recognition task and offers acceptable performance, it might not offer the best results compared to other sets when signals with a bandwidth of 7 kHz (WB) or above are available. These signals contain a frequency range beyond NB telephone speech with additional speaker-discriminative content, which was found to be substantial in [170, 171]. One of the sections of this book (Sect. 6.2) is dedicated to the examination of whether a greater resolution of the filters in the filterbank to emphasise the higher frequencies instead of following the mel scale is advantageous for the speaker recognition performance.

2.4.3 Effects of Communication Channels on Automatic Speaker Recognition

A growing number of applications using ASV often require the transmission of the user's voice to perform the authentication remotely, for example, for retrieving account information from the bank over the phone or for telephone-based credit-card transactions. Channel impairments like bandwidth limitation and speech coding introduce different distortions in the original speech, which augments the non-desirable within-speaker variability. Besides, mismatch between enrolment and test utterances can be generated if these are transmitted through different communication channels, causing a decrease in speaker recognition performance. Next, investigations of how different degradations affect the automatic speaker recognition are reviewed, as well as some techniques for distortion compensation.

Although the effects of coded-decoded NB and WB speech have not been compared before for automatic speaker recognition—this is one of the purposes of this book—the advantages of the frequencies beyond NB have already been assessed for clean, unprocessed speech. The earliest studies revealing the importance the frequencies incorporated in WB were [103] and [220] for text-dependent and for text-independent speaker recognition, respectively. It was found in [276] that the speaker identification accuracy improves as the sampling rate increases, up to a

sampling frequency of 11,025 Hz. Sampling rates of 22,050 Hz and higher caused a decrease in performance. GMM-based experiments from uncoded speech suggested that speaker verification was more accurate for signals with 16 kHz sampling frequency in comparison to signals sampled at 8 kHz [22, 36, 131], also in the presence of background noise and of microphone mismatch [205]. The sub-band analyses [20, 115, 170, 195, 230, 274] in Table 2.1 indicated the relevance of frequencies beyond 4 kHz for speaker recognition.

The great majority of past studies have addressed the effects of NB (and only a few of WB) transmissions on the performance of different speaker recognition systems. The ASV performance was found to decrease with the codec bitrate [60, 177, 209, 243] since low bitrate implies a loss of information from the original speech. The performance decline was consistent with the decrease of perceptual quality [21, 209]. The work in [160] showed that A-law coding caused a lesser decrease in text-dependent ASV performance than GSM-FR coding and detected an improvement extending the NB bandwidth to 0–4 kHz. It was also noted that speaker verification is more affected than word recognition by NB coding [68, 209]. The modern Speex codec was examined in [249] where a strong relationship between automatic speaker recognition and compression levels was reported. The recent study in [179] addressed the effects of NB coding speech on i-vector-PLDA speaker verification and the differences in employing noise-robust feature sets. The best results were obtained when the PLDA was trained with speech of the same codec as the evaluation utterances, and it was observed that noise-robust features did not offer any improvement over MFCCs in this case. Only negligible effects of packet loss in NB for speaker identification and verification were found in [248] and in [21], where it was reported that packet loss affects the automatic speech recognition performance to a greater extent.

The investigations in [22] and in [93] showed that it is possible to perform speaker recognition experiments employing features extracted from speech encoded with GSM codecs (not audio signals), which improved the recognition from cepstral coefficients. A pseudo text-independent approach to perform speaker verification from parameters of the G.729 codec, matching the performance offered by MFCCs, was given in [191].

The alteration of phonemes caused by channel transmissions involving coding and decoding processes has been investigated for text-dependent ASV [202], phoneme recognition [141], and some forensic analyses [98]. These studies have reported the unforeseen alteration of formant frequencies [98, 202], and the alterations of consonants and fricatives in particular due to telephone and cellular channels with respect to clean speech [141]. However, only NB coding has been analysed. The sub-band studies of Table 2.1 employed either clean data or coded-decoded data in NB only but no WB codecs were applied. Besides, no comparison between clean and distorted data was attempted. Only [20] detected a decrease of performance between TIMIT and NTIMIT (its NB version), which was attributed to handset, bandwidth filtering and telephone distortions, yet no further explanation was given.

Not many studies have been found that address the effects of WB coding on speaker recognition. The influence of the AMR-WB codec was reported for forensic investigations in [42] and for biometric applications in mobile communications in

[71], where speaker verification was shown to be more accurate when the system employed the coded parameters compared to when it employed MFCC features. Only the work presented in [139] has been found that examined the effects of both NB and WB speech on speaker identification. The authors tested a HMM-based system—which is normally better suited for text-dependent speaker recognition—on a relatively small dataset of 10 speakers, and found no significant identification improvement with WB speech respect to NB.

The effects of mismatch originated by transmitting the utterances for enrolment and for test through channels presenting different characteristics have been addressed in several investigations for NB coding. The studies [60] and [59] showed the speaker verification accuracy under matched and mismatched conditions and that the application of a handset detector and H-norm normalisation or handset-dependent test-score normalisation (HT-norm) could improve the performance in all cases. The speaker verification results in [243] indicated that low bitrate coding in matched conditions caused poorer performance than higher bitrate in mismatched conditions. It was determined in [131] that the Speex codec in mode 8 (at a bitrate of 3.95 kbit/s) and the G.723.1 codecs were more suitable for creating speaker models, since testing with different versions of NB-transmitted speech provided lower EERs than models created from speech processed with other codecs. The mismatch caused by different types of microphones was analysed in [216] and in [227].

Various techniques to reduce the channel variations or the channel mismatch have been proposed in the literature, predominantly at the feature level, at the model level, and at the score level. Feature warping [197] and mel-cepstral feature set with Cepstral Mean Subtraction (CMS) are successful approaches for feature normalisation, although the latter can be improved by the methods proposed in [192, 275]. Transforming the speaker models is achievable with no *a priori* knowledge about the origin of the mismatch [192]. The typical techniques for score normalisation are, as mentioned before, Z-norm, H-norm and T-norm [9]. The combination of feature warping and T-norm has been shown to be more effective than other standard normalisation techniques for cellular data [13]. The study in [177] revealed that the codec parameters can be useful to reduce the speaker verification error generated from NB coding. Background noises are challenging when speaker recognition services are to be used on handheld devices. The system should be adapted to expected environmental conditions by modelling the often unknown noise characteristics [184].

Also to mitigate the channel effects, the information for speaker authentication can be embedded in the speech transmission in the context of distributed speaker recognition using the ETSI Aurora standard [29, 92, 206], which was initially intended for distributed speech recognition. This approach is, however, not considered in this book. The present work rather focuses on a general scenario where the call receiver can be either a human or an automatic system (Fig. 2.1), and is not constrained to automatic speaker recognition on the mobile network as in the case of distributed speaker recognition.

2.4.4 NIST Speaker Recognition Evaluations

The NIST SRE challenges, already introduced in the review of human speaker recognition literature, have not only served to compare the performance of different speaker recognition systems under the same evaluation (enrolment/test) conditions; they have also enabled researchers to validate new approaches to foster advances in the field [57]. The task in the SREs is single-speaker detection or verification from provided segments of conversational speech for enrolment and for test. Same-gender trials of different conditions are proposed, generally varying on whether microphone or telephone recordings are to be compared with or without mismatch, on their lengths, and on other recording conditions. For each trial, participants must submit the score and/or the true/false decision output by their speaker recognition systems [96, 174].

The JFA system, modelling both speaker characteristics and channel effects, has been shown to outperform standard methods like GMM-UBM dealing with the SRE cross-channel conditions. Later, it was shown that the i-vector used as features and a simple classifier produced better results than JFA on the 2008 SRE [53, 151]. The i-vector approach combined with PLDA compensation, offering excellent discriminative capacity and small dimensionality, is the state-of-the art speaker recognition system employed in most current commercial and experimental applications [194]. Different techniques based on i-vectors are able to achieve less than 2 % EER[1,2] on the latest challenging NIST data, recorded under real telephone and microphone situations. PLDA is often combined with several other back-ends by applying fusion techniques in order to achieve the best performance depending on the nature of the given data for enrolment and test and on the testing paradigm [152, 229]. Current efforts are being conducted towards domain adaptation techniques,[3] that is, to counteract the effects of mismatched development data, which are often unlabelled [86].

Both the JFA and the i-vector-PLDA recognisers require large amounts of audio data for system development, obviously from speakers that do not appear in the evaluation test. Most researchers working with these approaches chose to develop their systems with combined data from NIST challenges and from the Switchboard corpus [100, 144, 149, 229]. It is required that the development set cover a variety of channel conditions that are expected to be encountered at verification time [149, 179]. Because mostly each speaker is recorded over only one phone channel, it is sometimes difficult to choose appropriate training data for NIST evaluations [148]. Regarding signal bandwidth, the data from the NIST challenges were, until the SRE 2012, band-limited to 300–3,400 Hz or clean speech sampled at 8 kHz [96]. For the telephone conditions, the audio signals were previously transmitted through

[1] The results of the NIST SRE 2012 challenge are reported in http://www.nist.gov/itl/iad/mig/sre12results.cfm, last accessed 28th September 2014.

[2] Already computed i-vectors were provided in the NIST 2014 Machine Learning Challenge with the aim of involving the machine learning community in the speaker recognition task [94].

[3] The Domain Adaptation Challenge (DAC) was organised in the summer 2013 by the Johns Hopkins University (JHU). The challenge description is given in http://www.clsp.jhu.edu/user_uploads/workshops/ws13/DAC_description_v2.pdf, last accessed 28th September 2014.

NB channels employing μ-law coding. The microphone (not transmitted) data were provided with a sampling frequency of 8 kHz to match that of the telephone data. Differently, in the SRE 2012, the released microphone data were sampled at 16 kHz. However, the participants, concerned with the challenging noisy and short-duration conditions, did not attempt to take advantage of the extended bandwidth. Besides, there is a lack of sufficient development data sampled at 16 kHz. Thus, the authors of the SRE 2012 submissions downsampled the 16 kHz microphone signals to 8 kHz and pooled it together with the telephone speech [78, 229, 255, 272].

The authors of [36] had access to the originally recorded conversational data from NIST SRE 2005 and SRE 2006 (microphone speech with 48 kHz sampling frequency) and to the SRE 2010 microphone data sampled at 16 kHz. This enabled the study of the effects of sampling frequency on the speaker verification performance, based on i-vectors and on inner product discriminant functions (IPDFs) [35]. Different systems, using either 8 or 16 kHz signals, were trained and developed on SRE 2005 and 2006 data and evaluated on part of the SRE 2010 data with the corresponding sampling frequency. The results indicated a considerable increase of performance when extending the speech bandwidth.

Some of the remaining challenges for ASV technologies are: achieving acceptable error rates with minimal amount of enrol and test speech, improving robustness against human or technical voice impersonation, and reliably assessing the systems performance accounting for their requirements in real use.

2.4.5 Comparison Between the Human and the Automatic Speaker Recognition Performance

The performances of listeners and of automatic systems recognising speakers have been compared in previous studies with the aim of identifying sources of errors and of finding improved features for automatic systems. It is also interesting to compare both performances in forensic settings, where the human and the automatic decisions can be fused to achieve the best performance, as proposed by the NIST HASR challenges [240]. Various analyses have indicated that the performance of automatic speaker recognition is higher than that of human listeners for non-degraded speech even if voices are disguised [142], while humans tend to outperform automatic systems in the presence of mismatch introduced by different communication networks and background noise [3], and handset variation [236]. However, the comparisons between the human and the automatic capabilities are highly dependent on the datasets employed for system development, on the degree of the system adaptation to channel variability or to noisy conditions, and on the listeners' expertise. More recently, it has been shown that the JFA system evaluated on a subset of the NIST 2008 data yielded 20 % EER whereas the fusion of human decisions produced 22 % EER [102].

An accurate comparison between human and machine capabilities under exactly the same training conditions is not possible. The paradigm of speaker verification, in which the performance of automatic systems is evaluated, differs from human speaker identification, which is more natural when listeners recognise interlocutors. Furthermore, humans and machines learn the speaker voices through different processes which differ in amount of data and audio content, i.e. human familiarisation versus system development followed by speaker model training from the enrolment material. Human memory and listener's inattention or tiredness throughout the listening test are additional factors affecting the human performance which are not pertinent to machines. In this book, the comparison of speaker recognition performances is not attempted as such, although relations between human identification accuracy and automatic verification scores depending on the degree of signal degradation are presented.

2.4.6 Literature on Automatic Speaker Recognition and This Book

The reviewed studies have indicated that the signal bandwidth is crucial for correct speaker recognition since the speaker-specific information is concentrated at different regions of the voice spectrum. The frequencies beyond the range of NB communications are effective for speaker recognition, although the extent of the benefits of WB over NB communications has not been carefully examined before. Large multisession data corpora from the NIST SRE evaluations are publicly available yet only a small set with sample rate of 16 kHz, released in 2012, would permit to experiment with an extended bandwidth, which has so far been overlooked in the literature.

In this book, the performance of the standard systems GMM-UBM, JFA, and i-vector is assessed under the effects of various channel distortions in different bandwidths, employing the common MFCC and LFCC features. This work does not intend to develop an improved speaker recognition system that would outperform the existing ones. Its objectives are rather to determine and compare the extent of the performance degradation caused by different speech transmissions, commonly encountered in biometric applications and in forensic investigations nowadays. Besides, the discrimination ability of different frequency regions and of certain phoneme classes under NB and WB coding is examined. The verification performance is related to instrumental signal quality measurements and to the human speaker recognition rates.

http://www.springer.com/978-981-287-726-0

Human and Automatic Speaker Recognition over Telecommunication Channels

Fernández Gallardo, L.

2016, XII, 169 p., Hardcover
ISBN: 978-981-287-726-0